The Problem of History in the Age of Abundance

By Ian Milligan | December 11, 2016

Our collective cultural heritage faces a serious problem: In the digital age, we can document and "remember" more than ever before. But the scale of historical material is so huge that it will decisively change how scholars interested in the past research and write. We are not ready.

Here's just one example of the challenge: a GeoCities personal home page from 1996. Founded in 1994 by David Bohnett and John Rezner, and closed in 2009 by the Yahoo corporation, GeoCities provided free websites to anybody who wanted one. Users could enter their email address and receive a free megabyte (eventually two, then 10) to stake their own space on the growing information superhighway- be it a Buffy the Vampire Slayer fan site, a celebration of a favorite sports team, a family tree, even a child's tribute to Winnie-the-Pooh.

People took to GeoCities with a passion: By October 1995, the first 10,000 users had created sites; two years later, that number had reached one million. By 2009, seven million users had accounts on GeoCities. The archived copy of the site we are left with today contains about 186 million distinct URLs. But if you type "GeoCities.com" into your web browser, you will see only an advertisement for Yahoo's website-hosting service. To see the site in all its glory requires a trip to the Internet Archive's Wayback Machine, which collects pages through a series of automated web crawls, overseen by archivists at their headquarters in San Francisco.

Historically, ordinary people did not leave behind many records, forcing historians to learn about them from the scant moments when they came into contact with large record-keeping institutions like censuses,churches,poor rolls,or the criminal-justice system. Between 1674 and 1913, the Old Bailey, the central London legal court, collected the transcripts of 197,745 trials. Today's Old Bailey website describes its holdings as the "largest body of records documenting the lives of non-elite people ever published." This is not hyperbole. For 239 years, 197,745 trials are a good standard of historical documentation. Compare that with the seven million users and 186 million "documents" that were generated on GeoCities in 15 years, and which represent only a fraction of the web of that era, and you get a sense of the enormous scale historians now confront.

The online record is not entirely representative of society; considerable barriers to web access and publishing still exist along lines of race, ethnicity, class, and gender. Still, events, feelings, and ideas are now being recorded that would never before have been, and they are being left behind by the sorts of people who used to be largely absent from the historical record.

Consider the intellectual implications. Much scholarship covering periods after 1996, the year widespread web archiving began at the Internet Archive and several national libraries around the world, will only be credible if it incorporates this born-digital information. Imagine writing a history of Bill Clinton's scandals during the mid-1990s or of the September 11, 2001, terrorist attacks without using archived websites. Imagine approaching the subject of the Iraq War without taking into account the posts and thoughts of deployed soldiers as they played out across the web. The same goes for any number of social and cultural topics, from Michael Jackson to Tamagotchis. It would be intellectually dishonest to tackle those topics without turning to the web. Yet nobody can read every source document available, or even a significant fraction of them.

With sources like the Old Bailey, historians have to scrounge for scraps to reconstruct a narrative past. Now, instead of too little information, we have too much. The late Roy Rosenzweig identified this predicament as early as 2003 in a pivotal American Historical Review article, but we are only now beginning to see its true scope.

Some of the challenges are technical. The Internet Archive contains more than 445 billion web pages, which in August 2014 consisted of more than 10 petabytes of storage (a petabyte is 1,000 terabytes). The numbers are much higher today. We need to tum to our colleagues in computer science and information faculties to find ways of making sense of this information deluge. Right now, web-archive users generally need to know the URL they are searching for and have access to a Wayback Machine (or similar interfaces run by national libraries like the Llbrary of Congress). Cognizant of this limitation, the Internet Archive and several other web archives around the world are working on a new search engine for their collections.

Few of us realize the skewing effect search engines have on our lives. Because of the algorithms that determine the order in which results appear, we are infinitely more likely to read a result on the first page of, for example, a Google search than if it appeared on the hundredth or thousandth page. The same goes for the tools we use to explore born-digital historical sources. If, when I am writing a history of how everyday Americans responded to the Monica Lewinsky scandal, I run a search in the web archive and read and use only the first thousand results to my query, the history will have almost been co-written by the search engine. It decided what results I would see, and which I would not.

Relevance-ranking algorithms will be an unavoidable part of research in the digital age. The Big UK Domain Data for the Arts and Humanities project, which explored use of Britain's web archive, developed a search engine that eschewed relevance ranking for fear of creating a "black box" (for much the same reasons discussed above), instead just displaying documents in order of their collection date. Yet this decision brings its own risks, making it akin to conducting historical research via Twitter stream, as potentially dangerous as relying on search-engine algorithms one does not understand.

In other words, historians cannot uncritically depend on ranked keyword search results. They need to help develop, craft, and make sense of these new and emerging digital tools.

My colleagues and I at the University of Waterloo, the University of Toronto, and Rutgers University at New Brunswick have realized this with a series of "datathons" held over the past year at the University of Toronto and the Llbrary of Congress. Interdisciplinary teams were formed both to "yack" - talk about the problems facing our cultural heritage - as well as to "hack" together particular research results or applications (such as a way to extract content of particular interest from web archives). Just as importantly, we need to bring disciplines together: Developers need historians to provide input, and historians need developers to help realize their humanistic goals. While the boundaries are fluid - programming historians and humanistic computer scientists - this form of interdisciplinary collaboration suggests to me a meaningful way forward.

The challenges are not just intellectual and technical, but also ethical. The overwhelming majority of web pages collected on Geo Cities were created by people who were unaware that their information would live on within the Internet Archive or other libraries. They did not consent to have their sites included in these repositories; nor did many have access to the robots.txt file that would exclude their sites from crawls such as those performed by the Internet Archive. We face the same issues today when working with people's social-media streams. Just because a Twitter account is public does not mean that it is necessarily ethical to embed, quote, and reproduce content without consent or consideration.

Yet given the short life span of the average website-the oft-bandied-about figure is around 100 days - it is important to collect the material before it disappears. And there really is no better way to collect it. Attempts by institutions such as the British Library to garner websites through an opt-in process have produced extremely low response rates (would you answer the sketchy email from a web archive that landed in your spam filter?). The opt-in method is an approach that would lead again to a disproportionate archival presence by powerful voices, corporations, and others who may be more likely to agree to seeing their records included.

The ethical onus, therefore, must lie on researchers. In many respects, we treat websites as akin to publications. Given the public-facing nature of the medium, this is not unreasonable. Strictly speaking, it is legal to quote from websites and biogs, as well as from social-media streams like tweets. Legal, of course, is not a synonym for ethical, as a host of commentators, academics, and activists have noted. Contemporary research into online behavior, such as Danah Boyd's study of online teens, reveals complicated attitudes toward privacy, social norms, and etiquette. As one student explained to Boyd, "I wouldn't go to my teacher's page and look at their stuff, so why should they go on mine to look at my stuff?"

Oral historians, who have long operated outside of formal archives and instead in the living rooms and workplaces of their subjects, may offer a useful model for scholars using web archives. Recognizing the implications of working with living people, we subject oral historians to considerable professional and institutional oversight. In Canada, where I work, they must complete an online course and quiz on the treatment of human subjects, and all oral-history projects must be proposed and overseen by an institutional review board. American scholars take similar precautions. The precise role IRBs should play in overseeing oral-history projects is sometimes debated, but there is general agreement that external regulation is valuable.

Potential considerations for working in the web age of historiography include scale (the larger the number of websites, the less fraught it might be when we quote from individual pages) or expectations of privacy (a famous site with thousands of visitors has a different expectation of privacy than a private, personal home page tucked away in a corner of the web).

Ultimately, web archives offer power: the power to reconstruct online lives, to peer into the minds and thoughts of millions of people from 20 years ago, and to move toward a potentially more democratic form of history. We need to think about the role of algorithms, to reach out to our librarian and archivist colleagues, and to begin a broader conversation about the implications. Without taking these steps, we will not be able to write honest histories of the 1990s or beyond.

Ian Milligan is an assistant professor of Canadian and digital history at the University of Waterloo and a 2016-17 fellow at the University of Toronto.